
Nr.: FIN-003-2020

# Ezennaya-Gomez - Rethinking Privacy-Knowledge Modeling: About Uncovering Accepted Data Collection Business Practices as Privacy Risks

Salatiel Ezennaya-Gomez

Arbeitsgruppe Multimedia and Security

Technical report


Fakultät für Informatik
Otto-von-Guericke-Universität Magdeburg

Nr.: FIN-03-2020

# Ezennaya-Gomez - Rethinking Privacy-Knowledge Modeling: About Uncovering Accepted Data Collection Business Practices as Privacy Risks

Salatiel Ezennaya-Gomez

Arbeitsgruppe Multimedia and Security

Technical report (Internet)
Elektronische Zeitschriftenreihe
der Fakultät für Informatik
der Otto-von-Guericke-Universität Magdeburg
ISSN 1869-5078

Fakultät für Informatik
Otto-von-Guericke-Universität Magdeburg

**Impressum** (§ 5 TMG)

*Herausgeber:*
Otto-von-Guericke-Universität Magdeburg
Fakultät für Informatik
Der Dekan

*Verantwortlich für diese Ausgabe:*
Otto-von-Guericke-Universität Magdeburg
Fakultät für Informatik
Salatiel Ezennaya-Gomez
Postfach 4120
39016 Magdeburg
E-Mail: salatiel.ezennaya@ovgu.de

http://www.cs.uni-magdeburg.de/Technical_reports.html

Technical report (Internet)
ISSN 1869-5078

*Redaktionsschluss*: 28.04.2020

*Bezug:* Otto-von-Guericke-Universität Magdeburg
Fakultät für Informatik
Dekanat

# Rethinking Privacy-Knowledge Modeling: About Uncovering Accepted Data Collection Business Practices as Privacy Risks

Salatiel Ezennaya-Gomez

Multimedia and Security Lab (AMSL)
Otto-von-Guericke-University Magdeburg
salatiel.ezennaya@ovgu.de

**Abstract.** Within some new law frameworks such as General Data Protection Regulation (GDPR), and California Consumer Privacy Act (CCPA), the consent of users is required for processing their sensitive data. This situation presents a motley collection of ways to provide user's rights over the collected data. To understand the legal implications of consent, the current Informed Consent (IC) landscape and its implications, it is necessary to understand the rationale of the business models behind generated data collections. In this paper we motivate needs on privacy-knowledge models proposed in the literature. Moreover, we intend to identify challenges that involve the mobile landscape (i.e. the nature of applications and data collection) related to privacy. This work does not intend to be a systematic review of the literature in threats to privacy models, rather to provide insights in the diverse approaches of interpreting informational privacy requirements to achieve user-centric and self-determination privacy management in system design for the field of mobile devices. In this paper, we describe the data collection business model as a dynamic system by bringing into focus the need to rethink the current practices in this field, which in our opinion, poses risks to data owners as well as processors.

**Keywords:** Informed consent, Threat modeling, Privacy-by-Design, Data scandals Mobile devices, Multiparty privacy.

## 1 Introduction

More and more scandals on misuse of sensitive data come to the spotlight. Data scandals such as Clearview AI [1] with its face recognition app, or Zoom Video, a company that offers a free online video conference system with security problems, and non-transparent practices on data processing [2, 3]. It has skyrocketed in stock market more than 100% from the beginning of 2020 due to the uncertain situation for the virus COVID-19, and lock-down measures that most of the countries around the world put in practice [4]. These non-transparent practices rise the need on studies about well-accepted data collection practices by the data industry, that may entail privacy risks to data subjects, and therefore,

the society and the future of privacy as it is understood. This is the reason of privacy concerns lately these years about the data exploitation economy, and that is what many organisations world-wide and privacy experts are denouncing it based on facts such as data breaches and data scandals [5, 6].

Drawing an overall picture of the current situation for data collection, the data economy can be represented as a balance, in which on one side there is the data subject who is the provider of the good, and on the other side, there is the data industry, which creates and offers tools and services made from this good. In one hand, sharing data and exploiting data is beneficiary for both parts. One receives a more individualise services while the other make profit of it. However, information is power, and in the presented situation one of the sides hold information, thus there is non-equilibrium in the balance. Furthermore, instead of one data subject we find cases (applications) where the data belongs to more than one subject, we call this multiparty privacy aspects [7]. In this paper we mainly focus on one subject to summarize the basic findings.

From the technical point of view, machines present to be a solution to balance this situation providing tools to inform, to enforce user privacy preferences, to interpret policies, and to assess at what extent the privacy policies satisfy user preferences. And last but not least, these systems should communicate to the data subject in case of both, security and privacy vulnerabilities regarding the legal, security and privacy protection context that the dataset is acquired. Trust systems that log and analyses the industrial context and practices behind doors in the industry side regarding data flow, may be a way to provide real transparency to the user at the time to accept some practices. The main goal in system design is to understand and model the informational privacy requirements, and threats associated to design trust systems that automatically follow the Privacy-by-Design principles [8].

For the aforementioned reasons, models of privacy knowledge and risks are an important key to understand privacy requirements and factors on decision-making, which may influence the user while giving consent or accepting privacy policies, along with risks assessment and threats to privacy. That is, changes on the context of the data subject and the data consumers (i.e. business side) may provoke changes on data subject decisions regarding the outsourced data, that may affect the terms agreed of a given consent.

In this paper we motivate needs on privacy-knowledge models proposed in the literature for privacy system design. We presume that an understanding is required of the role of consent for the data economy from a security and privacy system design perspective. In fact, to understand the current motley privacy policies landscape and its implications, it is necessary to analyse data scandals and their controversies that data business collection model generates, which may affect the user's will to consent information disclosure. Thus, identifying challenges that involve the mobile landscape (i.e. the nature of applications, privacy policies and data collection practices) related to privacy, we conducted a non-exhaustive literature review, aiming to answer: What is the role of privacy policies in the data collection business model?; Is the data collection business context

included on those models?; What aspects of privacy models take into account data breaches as a threat to privacy/user's consent?; Are the proposed solutions suitable for the current use of biometric data for the smartphone application landscape?

The organisation of the paper is the following: Section **??** presents the problem understanding of data exploitation followed by a review of the literature in privacy modeling and applications. Subsequently, in section 4, an analysis of data breaches is conducted in which sensitive data, such as videos/face images which involves biometric data and biometric recognition techniques, has been compromised. Finally, in sections 5 and 6, we discuss the literature review findings in relation to data scandals, the mobile applications and biometric data usage.

## 2 Problem Understanding

### 2.1 Data Exploitation: The Business Model of Data Collection

Big Data Business models are usually based on Data-as-a-Service (DaaS). Such models could be classified into three main categories: Data collection by developing digital products, e.g. collected data by users' interaction with a platform. Machine Learning as a service, i.e. algorithm application to analyse aggregated data. And finally, data monetization, that is selling access to information for diverse goals, but mainly advertising purposes targeting individual people [9].

Data collection business can be identified as a complex dynamic system. As new actors, roles, activities and industrial agreements take place, the data market situation is constantly changing. To conduct data collection, the purposes should be reflected in the privacy policies as some the protection laws require. That is, the data subject should be informed about the data collection purposes, along with data sharing practices (e.g. purposes of collection and storage, third parties with whom datasets are shared) and be able to revoke this agreement. A close concept to this idea is the dynamic consent [10], whilst context around datasets changes, conditions for data processing also must change.

A global scenario —in this case, based on GDPR elements [11], though it is applicable for any law framework such as CCPA [12]— which is identified as a distributed system. An overview of the global scenario is shown in Figure 1. The elements are also described in some legal ontologies based on GDPR [13–16]. There are entities who perform actions (i.e. data processing, dataset storage, data sharing, permission request, purpose definition etc.) according to their role defined by business processes. Adopting GDPR definitions, the roles are: Data Subject, Data Controller, Data Recipient, Data Protection Officer (DPO), and Data Processor. Each actor, except the data subject, is represented as a repository which may store a data subject's dataset, and interacts with other actors' repositories within their business processes model. Within this model, we identify, up to four, information exchange channels described as interactions:

– **Data flow**: The channel where data is transferred from one entity to another. A dataset is shared across several repositories. Each repository holds an

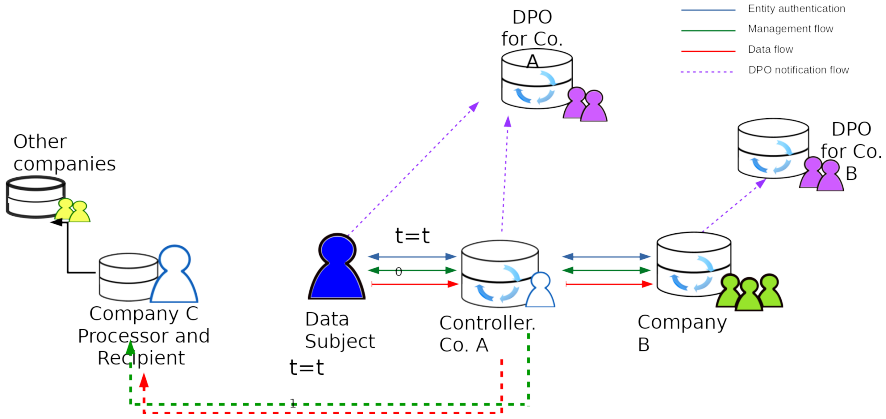


**Fig. 1.** An example of the management flow and data flow: Let Company A be offering a service via mobile app, in which it is necessary to collect sensitive data, e.g. face and voice. This company has an service-level-agreement (SLA) with Company B, which stores and processes the datasets for Company A. At time $t = t_0$, Company A creates an electronic consent (e-consent) for the data subject who agrees the processing of his/hers dataset. Later, at time $t = t_1$, Company A creates new agreements with another Company C to carry out their business, selling their solutions on processing data. As new actors come into the scenario, their actions may affect the current agreement between the data subject and Company A.

instance of this dataset. Each instance is a derivative of the original dataset. Formally, the formula can be described as follows: $d^* = f_2(d' = f_1(d, P_1), P_2)$ , where $d$ is the original dataset, functions $f_1$, $f_2$ are transformation functions (e.g. pseudonymisation operations), and $P_1$, $P_2$ sets of policies applied to $d$ and $d'$, respectively.

- **Entity Authentication flow**: Each stage in which an entity is authenticated to another entity as a previous step of sharing any resource. Authentication mechanisms are well implemented in current security systems.
- **Management or Consent Agreement flow**: The information flow corresponding to privacy policies and consent among the actors. This flow is the focus of our study.
- **DPO Notification flow**: Some actors, such as the controller, should notify the Data Protection Officer (DPO) of the current situation of consent, requests, and conditions. In this case, the DPO is not acting as a Trusted Authority, but as an passive actor auditing the consent and processes.

The ideal situation is setting an equilibrium between parties. Within this exchange context, the value of a dataset to the data subject is balanced with the value of that dataset to the aggregated data processor within its domain. Therefore, the service or tool value offered is proportional to the value of the dataset exchanged along with the related costs. An important factor that aims

this equilibrium are the data regulations, which are forcing companies to enforce transparency, and accountability regarding customers' privacy. For this purpose, they relay on data subject consent and privacy policies. Hence, privacy policies and the user's consent are static entities that determines the relationship, actions and conditions between the actors of the dynamic system.

### 2.2 Privacy Policies, Biometric Data and Mobile Devices

Biometric data is a special category of data according to GDPR. Acquisition is allowed, if and only if, there is consent (always meeting freely-given, specific, informed, and unambiguous). Biometric data should be treated as medical data in a way, while consent has the same features as consent in the medical field. However, there are applications available in market places (e.g. Google Play) which for its functioning biometric data is mandatory, such as application for unlocking phone with face and/or voice in form of video/audio, for instance. In such cases, consent is given through the acceptance of the privacy policies presented at the installation stage of the app. The acceptance of these policies by the user, implies the disclosure of sensitive data, including to third parties.

In the mobile ecosystem, privacy policies, and terms of use are the key to get access to user's data. Privacy policies discloses the data usage, data sharing and data processing conditions performed by the data controller. However, it has been proved that privacy policies in the mobile ecosystem do not represent the reality of these policies and practices.

In the last five years, research studies on mobile devices has been conducted about decision-making on privacy preferences and rewards [17], mobile apps traffic analysis, and mobile app permissions on privacy policies [18]. They have highlighted inconsistencies between mobile apps privacy policies and apps code behaviour [19, 20], showing an unbalanced situation, where the data industry's revenue for capitalizing data with detriment to data subject's benefit.

The current situation in the mobile landscape is such that data subject has to rely on the privacy conditions and notices that each data holder provides. Besides, the user depends on the level of transparency that the data holder is willing to disclose.

## 3 Research Efforts on Modeling and Applications

To tackle the complex problem of privacy, many approaches in many ways have been published lately in the literature. Since data regulations came into force, the effort is focused on creating systems and technical mechanisms capable to interpret and comply with these regulations. From architectures to ontologies, the common goal is to understand and model the informational privacy requirements, and threats associated for privacy-knowledge modeling, and therefore aim to design trust systems that automatically follow the Privacy-by-Design principles. We reviewed in the literature these three interconnected topics: systems which ensure consent, context-aware systems for privacy and consent, and ontologies

on privacy-knowledge and threats and provide an overview in the remaining part of this section.

***Ensuring Consent*** Informed Consent is in focus as a building block in privacy system design, since obtaining consent for data processing is mandatory for some categorisation of data(e.g. PII, medical data, biometric data). Foundations for online informed consent have been published concerning a more defined design framework long before these data regulations [21, 22]. Moreover, there are solutions underway such as consent templates, to provide a tool to manage all signed consent forms and registered by the user [23, 24].

Concurrently, to ensure that machines "understand" data regulations implications, standardisation groups are working on modeling machine-readable languages to describe in form of policies the current data regulations, such as the IEEE P7012 Standard for Machine Readable Personal Privacy Terms[1]. Also, there are other works on the topic. Robol et al. have proposed a modelling language for consent requirements compliant with GDPR in the medical field [25].

Related work on architectures which ensure given consent were presented before data regulations came into force. Within the project EnCoRe, Mont et al. presented an architecture which enforces consent and its revocation using sticky policies, [26]. In line with Mont el al.'s work, Pearson's work is very extensive in privacy system design for Cloud using sticky policies [27, 28]. Sticky policies has been proposed for Cloud Computing mainly as an access control, and for identity-based encryption, [29, 30]. Other architectures have been proposed as personalised privacy assistants for the IoT [31]. Generally, the proposed architectures in the literature, rely on a trust third-party or authorities (e.g. in the cloud), where the user manages the outsourced data (e.g. who has access to the dataset).

***Context-aware Systems*** Context-aware systems have been studied for decades. Context is information that can be used to characterise the situation of an entity. Context-aware applications look at the 'who, where, when, and what' of entities, and use this information to determine a situation [32].

In the field of privacy and consent, there are some ideas that coincide in some points. Pearson et al. propose a decision based support system assessing context for privacy requirements [33]. They present an approach for design pattern selection based on context. The presented system has several actors: Context Handler, Context Processing Rules, Context Processing Rules combination (CPRC); and selection Criteria generated by the CPRC, design pattern selection rules, and a design pattern selection rules combination. They provide the idea of a Context Handler that quantifies this information input. In the proposal, contexts are pre-defined by the system administrator for the pattern selection according to the following points: sensitivity of data, location of the data, potential location of transferred data, sector, number of users of the system, whether anonymous

[1] P7012 – Standard for Machine Readable Personal Privacy Terms – `https://standards.ieee.org/project/7012.html`

data set could be usable, contractual restrictions, cultural expectations, users role in the organisation, security deployed in the infrastructure, and intend of system design.

A similar concept for context within a definition of consent ontology, providing a consent and data management model was presented by Fatema et al. [13]. Within the model, they note that consent can be influence by context or situations: data breaches, changes in purposes, modification or withdraw of consent, expiry data of consent or dataset, external context such as business agreements. These factors influence lifecycle of Consent which also rules the data lifecycle. However, neither Peason et al. nor Fatema et al. specify the extend to which these factors affect, what sources are needed to retrieve the context information, or how they are/can be modelled.

***On Privacy-Knowledge and Threat Detection Ontologies*** Ontologies are the formalisation of terms in a domain and the relations among them [34]. In the last four years, ontologies gathering privacy requirements and GDPR have been published. Table 3 summarises the following discussion by listing its primary application domain, applied ontological methodology, and factors included. As previously mentioned, in 2017, Fatema et.al proposed an ontology, an extension of Provenance Ontology (PROV-O), that formalises a generic model for the notion of consent by identifying and describing the important concepts and relations [13]. The basic purpose of giving consent is providing permission to perform personal data processing for specified purposes. This work is inline with the architectures and machine-readable languages previously mentioned. Pandit et al. continues the work done by Fatema et al. proposing the Gconsent Ontology, based on GDPR [16]. According to them, there is a lack of the existing work which is only focused on the given consent, but none is focused on aspects such as other states of consent (e.g. not given or refused). Moreover, similar proposals on legal basis were presented by Palmiriani et al., Geko et al., and Loukil et al. [15, 14, 35].

Palmiriani et al. presented an ontology GDPR-based well granulated in terms of modules that compose consent as specified in the GDPR. These modules are interconnected, which are: data and documents, processing (processing and workflow of data), purposes (based on legal basis: deontic formula and legal rules), agents, and rights. These last two modules are interconnected, which are actors and roles, and isolated from the firsts. In their work, they do not explicitly define context as an entity in the ontology, although context would be defined as a group of connected classes and subclasses. For instance, the personal data processing is valid within an Interval of Time and context (place and jurisdiction), as GDPR stresses where the consent is taken. Nevertheless, this is the only example of context that we could extract from this ontology.

Geko et al. proposed an ontology GDPR-based mapping Information Security (i.e. ISO 2700x) and the data regulation. With the three protection goals introduced in the Standard Data Protection Model (SDM), the ontology aims to express dependency between Obligation class and its subclass Data Security. For instance, some of these obligations defined in the regulation are: record of

| Ontology Approach | Primary Application | Methodology | Factors identified | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Actors Description | Retention Period | Collection Purpose | Data Shared with3rd Parties | Consent | Category of Data | Status of Consent | Context-awareness | Privacy Policy Violation | Device elements | System description | Information Security elements | Risk terms |
| Arruda et al.[36] | IoT | 101 [34] | | x | x | x | | | | | x | x | x | x | |
| Loukil et al.[35] | IoT (legal compliant) | extended version | | x | x | x | x | | | | | | | | |
| Geko et al.[14] | GDPR-based | 101 [34] | x | x | | | x | | | | | | | x | |
| Pandit et al.[16] | GDPR-based | 101 [34] | x | | x | | x | x | x | | | | | | |
| Fatema et al. [13] | GDPR-based | extended version | x | x | x | | x | x | | x | | | | | |
| Palmiriani et al.[15] | GDPR-based | MELON[15] | x | x | x | x | x | x | | | x | | | | |
| Gharib et al. [37] | Privacy | SLR(a) | x | x | x | | | | | x | | | | | |
| Haynes [38] | Online Privacy Risk | Own(b) | x | | | | | | | | | | | x | x |
| Li et al. [39] | Data Sharing | Grüninger [39] | x | | x | x | x | | x | | x | | | | |

**Table 1.** (a) Systematic Literature Review. (b) Step 1: Use Cases Step 2: Literature analysis Step 3: interview with subject experts. (c) Step 1: Describe Motivating Scenario, Step 2: Determine Competency Questions, Step 3: Derive Concepts and Relations, Step 4: Evaluate the Ontology.

processing activities, performing a privacy impact assessment, and processing must be compliant with the regulation and additionally with codes of conduct, [14]. In the exercise, data security is divided into two major classes: measures and properties. Specifically, the class 'measures' is divided into 'organisational' (e.g access control, audits) and 'technical' (e.g. encryption).

In security and privacy system design, privacy and security threat modeling are the hot topics in the literature. Threat modelling is a process that aids understanding of the possible attacks to a secure system and its assets. From the Privacy-by-Design principles point of view and under this definition, threat modelling is focused on processes/actions that may harm the data owner's privacy. Unlike in security threat modelling, threat identification methodologies for privacy are not that well researched, since there are few threat modelling methods for privacy in the literature [40]. Nevertheless, both topics, privacy and security, share the same principles accepted by the industry, that is to identify and analyse potential attacks. Regarding threat modelling methodologies, there is not an established method, but rather categories of methods (e.g. theoretical,

empirical, and not specified [40]) to achieve the same goal. Among these, and a suitable method for our goal is the LINDDUN methodology [41, 42]. It is a privacy threat modelling methodology which presents a breakdown of threats to privacy into categories and their relation to already categorised security threats, STRIDE ([41, 40]). LINDDUN presents a novelty of privacy threat categories: Non-compliance, and Unawareness. However, its major drawback is automatisation. The methodology for privacy threat modeling is entirely manual, and hard to automatise using machine-readable languages since it describes high level concepts for privacy threats.

During the design process, understanding individual's privacy expectations may be a complex task. Privacy-knowledge models have been modelled in the literature for different domains. In a non-exhaustive but insightful review done by Perera et al. about privacy-knowledge modeling for the IoT, they discussed the idea of open data markets [43]. In open data marketplaces, the data owner, and the data consumer (i.e. who is interested on the dataset), exchange datasets and interests prior to a negotiation. Perera et al. conclude the need on ontologies to capture knowledge on privacy expectations for the data owner. Therefore, we consider those publications on threats and risks ontologies related to privacy-knowledge modeling and data regulations.

Palmiriani et al. specified in their ontology the risk and riskiness degree of freedom and rights linked only to a type of data [15]. Gharib et al. describes an ontology based on a systematic literature review for privacy requirements [37]. The concepts of the ontology are organised into four main dimensions: organisational, risks, treatment and privacy. For risk and privacy dimensions, proposes concepts to capture risks that might endanger privacy needs at the social and organizational levels, and to capture the stakeholders' (actors) privacy requirements or needs concerning their personal information. Li et al. described a data sharing privacy ontology (DSAP) [39]. The DSAP ontology endorses transparency and accountability with respect to the personal data sharing when multiple participants are collaborating in a health research context.

Recently, Haynes presented an ontology of risk for online privacy. The ontology is being developed to allow flexibility in the definition relationships, and to adopt approaches used in the semantic web, [38]. Although, this ontology works the concept that an external and non-trusted attacker compromises user's privacy, this proposal is in line with our topic data and mobile devices. There are also ontologies for security threats considered in threat taxonomies, such as insider threat proposed by Costa et al., who developed an ontology on malicious insider threat indicator directed towards organisations [44].

One of the most active fields in terms of developing ontologies for the formalisation of privacy requirements, is the IoT application, in which context awareness, privacy preferences, and risks to privacy meet each other. Celdran et al. proposes an ontology and architecture, called SeCoMan (Semantic-Aware Policy Framework for Developing Privacy-Preserving and Context-Aware Smart Applications ) where context-aware policies are based on information extracted from location and user privacy preferences on sharing locations [45]. Moreover,

ontologies for legal compliant have been proposed by Loukil et al. [35]. Although, the authors coincide with Fatema et al on consent lifecycle, they present a detail version of an ontology capturing consent attributes (similar to those described by the GDPR), or privacy attributes as the authors named them.

Nevertheless, a thorough ontology for privacy protection and ensuring consent in the IoT was presented by Arruda et al. [36]. They based their ontology reuses knowledge from another semantic sensor network ontology called IoT-Lite. IoT-Priv is an extension of IoT-lite by adding privacy requirements.

While searching for privacy-knowledge models, there are plenty of proposed solutions on privacy-enhancing technologies, however, these solutions act as add-on for the current designed systems. They do not tackle the problem of designing machines that detects and prevent privacy threats as design secure systems do.

The closest applications to our aim are the IoT environment, and online risks analysis. The former, because a smartphone and/or gadgets can be part of it. Regarding the latter, the nature of mobile devices forces to have a transparent and mapped landscape on risks on data sharing and processing practices. A starting point towards the objective is the creation of an ontology capturing the privacy requirements, and the state-of-the-art threats along with potential risk use cases.

## 4 Data Scandals and Controversies under Scrutiny

To define an accurate privacy risk ontology for mobile devices, analysis of the existing real cases is necessary to know the nature of those risks. In this section, we described data scandals, questionable practices and controversial situations as examples of deep-rooted practices in the data-flow economy which present potential privacy risks. These real cases have a global privacy risk in common, that is misuse of sensitive data.

### 4.1 Use of $3^{rd}$ Party SDKs

In mobile application developing, the use of third party software developer kits (SDKs) is fairly common. An SDK provides the developers the integration of their apps with the SDK providers platform, and its features, e.g. ads, analytics. However, it has become another channel of real-time data acquisition for data brokers and such. Thus, it presents a challenge in informational privacy. Conducted traffic analysis studies demonstrated that some SDKs insert trackers, and moreover, they send data to third parties without data controllers knowledge [46].

The presented use case is about medical data and behavioural data sent to the SDK's provider (e.g. Facebook) while the app developer is acting as a middle agent. According to the analysis of the NGO Privacy International, some of the most downloaded applications for menstruation control in Google Play portal, have sent or send sensitive information to Facebook among other companies through their SDK and APIs (application programming interface). Applying

dynamic analysis in mobile applications, they have discovered when the apps user had introduced requested information by the app, this one sent it (e.g. mood status, if the user had nauseas or diarrhoea, the menstruation cycle). Furthermore, the app started sending information to Facebook, before the user gave the consent [46].

Surprisingly, the sensitive data sharing is very common among certain popular apps as similar studies have been confirmed, despite the law enforcement. However, when these studies have attracted attention on certain app companies, some have followed studies' recommendation changing also their sharing privacy policies. The results have detected data types, such as IMEI, Ad ID and GPS location, that provide linkability. These facts reveal that they are sharing data beyond the purposes of the app [47, 48]. Although in some occasions, such purposes are not well described or specific enough about this data collection and sharing purposes[2]. Threats detected in such cases are the following: Under Non-Compliance Threat (data sharing is performed before the agreement of the privacy policies), and ambiguous purposes —privacy policies are not clear and specific— [49, 42].

### 4.2 Cambridge Analytica and Facebook Data Scandal

The Cambridge Analytica (CA) scandal came to light in 2018, when a whistleblower revealed in the Guardian [3] how this company harvested user account information from FB to influence among the political elections, e.g. US elections, Brexit. However, the condemned occurrences happened before more restricted data protection regulations came into force It started in 2014, the social consequences caused by this company are being still evaluated in 2020 [4]. CA was a political data analytics firm which used "psychographic" models generated from social network data to target population. Basically, CA was a data broker company funded, only, on Facebook data. CA access to FB profile data of millions of users through an app developed by a company of a researcher (Dr.Kogan) in psychology of the Cambridge University.

When a FB user downloads the app offered by FB, the user accepts the privacy policies and give consent to the app to access user's personal information, implying also personal information from other connected people to this user in the social network. With this amount of data, CA would create tailored online content, e.g. websites, blogs, to target individuals belonging to a specific psychological profile.

[2] No Body's Business But Mine: How Menstruation Apps Are Sharing Your Data – https://privacyinternational.org/long-read/3196/no-bodys-business-mine-how-menstruation-apps-are-sharing-your-data

[3] A year-long investigation into Facebook, data, and influencing electrions in the digital age – https://www.theguardian.com/news/series/cambridge-analytica-files

[4] Fresh Cambridge Analytica leak shows global manipulation is out of control – https://www.theguardian.com/uk-news/2020/jan/04/cambridge-analytica-data-leak-global-election-manipulationJanuary2020

Describing the event in terms of GDPR elements, Facebook is a data controller and data recipient of data subjects datasets and FB profiles. From the other side of the diagram, Dr. Kogan's application is acting also as a controller, because is the one who is asking permissions to access to personal data. Finally, CA is the processor and recipient of the FB datasets. Considering that, a detailed description of the data sharing process carried by FB, Kogan and CA is not clearly provided by the parts, we assume that the sharing process started at the moment, the user consents and introduces his FB profile information. Therefore, FB grants access to CA to acquire all type of data that is in their datasets.

These facts present an example of abuse of information along with threats to the users privacy. Two major privacy design aspects are threatened: Transparency —due to the lack of notification to the data owner about data sharing and processing purposes— and Intervenability —since the user had no power to revoke any consent— [50].

Described threats fall down into a Legal category and Social Engineering (according to ENISA Threat Taxonomy [51]) with some extensions of the aforementioned taxonomies. The compromised point is localised mainly at FB and CA side, due to the fact that both entities did not supervise each other on data sharing and usage procedures. The risks identified are: Misuse and abuse of information and personal data, unauthorised activity (ML techniques application, such as profiling, with lack of knowledge by data owner and recipient), failure to meet contractual requirements (that is, laxative contractual terms for data sharing between FB and CA). Furthermore, between the controller CA and the data subject there was a clear no notification about the processing procedures and data acquisition.

### 4.3 Public Databases and Face Recognition

Clearview AI is a company which offers its face recognition tool to identify a person within their databases. Its major customers are law enforcement agencies which uses the tool as a crime-solving tool. The research tool is fairly simple to use. A picture of a target person is uploaded to Clearview AI database —which backbone is a database of millions of public pictures of people, mainly from social media— where face recognition is performed. The tool also provides links to websites where a public photograph of the target is available. Whilst there is no confirmed data about the what type of criminals are the most identified through this app, the application carries risks to privacy to anyone in public places [1].

This case, presents a huge controversial over the use of face recognition [4, 1, 52]. In one hand, any person can upload a photograph to a third party database (i.e. Clearview AI), in this case, law enforcement agencies are feeding the database with sensitive data. In the other hand, we have the default privacy settings of social media platforms. They are not blocking access (e.g. search engines, information retrieval algorithms) to those pictures of their users stored in

their servers. Moreover, this situation demonstrates that the privacy preferences are not Privacy-by-Default settings [8, 53].

## 5 Discussion

To tackle the problem of user unawareness, part of the solution is putting in focus actions carried on the business side regarding data processing and privacy policies. To bridge the gap between the dynamics of business models and the current static characteristics of privacy policies, we considered an evolution from current practices for privacy policies (as a static object) into a constant updating dynamic consent taking into account changes in context, such as data breaches, business agreements and current state of transparency of the companies. Firstly, to obtain a balance between the benefits of the different stakeholders in a data driven business model, and secondly to keep this balance intact over business case changes to minimise the privacy and business risks.

One particular situation in mobile devices is the type of data outsourced by the mobile device. The acquired dataset may be not tagged as sensitive, as described in some data regulations. That is, some protected attributes are not explicitly acquired for any specific purpose. However, in data exploitation, knowledge discovering is conducted through aggregated data, therefore, if a dataset seems meaningless, combined with other datasets (i.e. public or third-party datasets), can be significant in terms of discovery knowledge.

A high level modus operandi of the business model that may lead to data scandals is described in figure 2. It shows a global picture of the data exchange sequence among the different actors in the presented data breaches, except for the ClearView AI case, which is a concrete case of access to public sensitive data, such as pictures. When the user accepts the app's privacy policies, implicitly he/she is also accepting privacy policies of third parties (i.e. mostly advertising companies in cases that the app carries add trackers, e.g. data scandal case explained in subsection 4.1). Therefore, the user is giving consent to these third parties to access to phone's data, which may differ to the dataset that the controller has access to. Paths 1 to 3 shown in figure 2, usually are specifically explained in the privacy policies. However, the nature of the advertising business gives no option to unable third parties trackers embedded in the app code (path 4 in figure 2). Thus, such data flow through many different parties (generally unknown by the user) is a key threat to user's privacy. In addition, in such data flow diagram, access management can be enforced as a solution to control who is accessing to the dataset, as has been suggested in the literature. Nonetheless, this security solution does not ensure the application of machine learning techniques on datasets containing biometric data. As some studies have reported, there are risks during the retention period. We consider that another of the major risks is the machine learning application for biometric identification in the concrete case for mobile applications (unless the user has decided to allow it considering the risks). That is, whenever there is no purpose of authentication, and neither the application is not following the current biometric authentication standards.

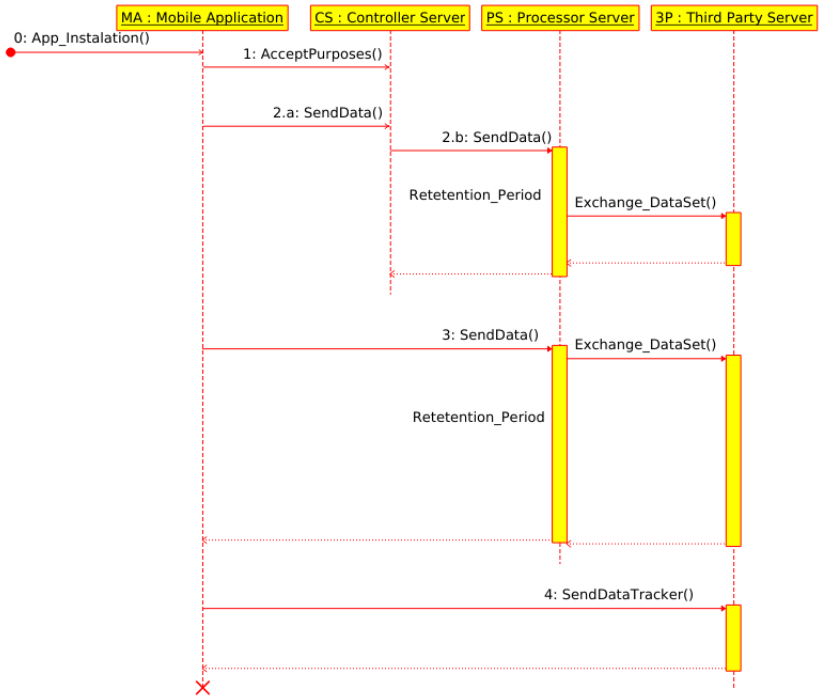


**Fig. 2.** A mobile application (MA) is installed (0), and the user accepts the privacy policies (1). Consecutively, the app sends the dataset (2.a) to the controller server (CS). Therefore, CS sends an instance of the dataset to the processor server (PS). Nevertheless, the app also can send the dataset (3) directly to PS, instead of CS. During the retention period, PS may share or monetise an instance of a dataset that may include MA dataset attributes. At some point, MA maybe uninstalled. Currently, applications do not notify the data subject any deletion of the original dataset, neither how many instances has been distributed among third parties. Path 4 represents the data flow from app's phone tracker to the tracker's company. By accepting CS privacy policies, third parties, that have their trackers embedded in the app's code by means of SDKs, have access to a dataset that may differ to the CS dataset. That is, the tracker may send information such as Advertising ID or non-precise location of the phone.

A lesson learned from the literature is the adoption of ontologies as a first step towards an implementation of privacy policies into machine-readable languages. In general, ontologies presented in the literature capture relations among privacy and law requirements. Among these ontologies, a few are defined on privacy risks or threats. However, those that do it, present a high level description on the topic. Therefore, to the best of our knowledge there is no ontology for mobile applications that defines relationships among privacy requirements, risks to privacy including privacy threats such as Non-compliance or Unawareness presented in

LINDDUN [42]. Nevertheless, these two categories fall short of modelling some situations, such as misuse of data even for a given consent to process the dataset. Moreover, most of the risks assessments are focused on organisations on information security risks, however few are on data subject awareness. This issue is discussed by Sion et al. in their privacy risk assessment [54]. They presented a calculation of risks for what they define "hard privacy" threats (i.e. privacy threats related to information security threats), although the model fails for "soft privacy" that are Non-compliance and Unawareness.

Another point of discussion in the literature is the application field. The evaluation of these ontologies has been performed in the medical, legal or research field. A drawback of these fields is that they are too restrictive compare to the mobile application landscape. Whilst in both fields, health and research, are very regulated fields for decades, the data mobile landscape has been recently regulated in terms of data protection. While in the former, there are plenty of proposed solutions for enforcing GDPR and given Consent (e.g. using sticky policies, etc.), the latter is far away from this controlled environment. For this reason, one of the mayor drawbacks of the proposed solutions are the flexibility and adaptability to the mobile landscape.

Furthermore, the majority of the proposals (e.g. architectures) assume two things: The data consumer's systems are honest which will stick to the rules of the privacy policies, and if there is an attacker, is a non-trusted malicious actor (either external or insider). Unfortunately, this is not even close to the reality of data exploitation, as many studies on application data flow has proved it, along with formerly described data scandals in section 4. The malicious actor is one (or more than one) of the members that takes part in the data flow. Then, is it possible to prevent malicious actions coming from one of the essential actors of the game? The question can be tricky to answer. However, one possible solution can be detecting modus operandi of the members, and evaluate them if they become a risk to the data subject interests, for instance as it is done for insider threat modeling [44]. For these reasons, we support negotiation stages before/during data life-cycle for an agreement on the data usage, as part of the concept of a dynamic consent introduced by Kaye et al. [10].

## 6 Conclusions

In conclusion, as highlighted in section 2 the data subject is at a disadvantage in data driven business models. This disadvantageous position, even if complying with current legal situation, is considered here as a potential privacy threat, creating privacy risks for the data subjects and business risks for the data collecting companies. While the users claim transparency from the industry, the latter claims guidelines on privacy system design as done for security system design beyond best practices, or law frameworks. To tackle the problem of lack of transparency in privacy policies, we consider an electronic consent concept as an exchange of interests between two entities, i.e. data holder and data subject, concerning personal identifiable information (PII) and sensitive data. A close

concept to this idea is the dynamic consent [10], whilst context around datasets changes, conditions for data processing also must change.

From the literature review in section 3, we identify the following main points to work on: the need of negotiation modules over data processing and sharing, accurate and transparent reports on real-time automated privacy risks assessment to the user done by the system. It could be done, studying privacy-knowledge models while associating them with privacy risks, that is, capturing and assessing all possible risks to the user (e.g. financial, legal, personal, and societal). Therefore, the benefits of data collection business would be in accordance to our idea of balance. In section 4, we have described real cases as business practices where data processing, disclosure of sensitive data under data regulation laws and vague privacy policies, may lead to risks for the user's privacy preferences and protection.

Bringing the paper to an end, detecting and profiling modus operandi of some business practices in data sharing/flow/privacy policies, and other interrelated factors of companies that have violated/presented a threat to user privacy, may be a step to integrate such modus operandi as risk for a further privacy calculus. As future insights, we propose further ontologies on mobile applications including privacy policies that affect data processing, close to the recent work done on online risks by [38].

## 7 Acknowledgments

The work presented has been supported in part by the European Commission through the MSCA-ITN-ETN - European Training Networks Programme under Project ID: 675087 ("AMBER - enhAnced Mobile BiomEtRics"). The author would like to thank Jana Dittmann, Claus Vielhauer, and Christian Kraetzer for their contributions on the presented ideas of the paper.

## References

1. TheNewYorkTimes, "The secretive company that might end privacy as we know it." https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html, 2020. Accessed 7 Apr 2020.
2. CNET, "Using zoom while working from home? here are the privacy risks to watch out for." https://www.cnet.com/news/using-zoom-while-working-from-home-here-are-the-privacy-risks-to-watch-out-for/, 2020. Accessed 7 Apr 2020.
3. B. Marczak and J. Scott-Rail, "Move fast and roll your own crypto a quick look at the confidentiality of zoom meetings." https://citizenlab.ca/2020/04/move-fast-roll-your-own-crypto-a-quick-look-at-the-confidentiality-of-zoom-meetings/, 2020. Acccessed 7 Apr 2020.
4. BusinessInsider, "Clearview ai scraped billions of photos from social media to build a facial recognition app that can id anyone here's everything you need to know about the mysterious company." https://www.businessinsider.com/what-is-clearview-ai-controversial-facial-recognition-startup-2020-3?r=DE&IR=T#then-in-march-another-new-york-times-piece-on-the-company-revealed-another-

stunning-detail-the-companys-founders-casually-gave-access-to-the-software-to-potential-investors-and-friends-who-immediately-abused-it-6, 2020. Accessed 7 Apr 2020.
5. C. Cadwalladr and E. Graham-Harrison, "Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach." https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election, 2018. Accessed 7 Apr 2020.
6. TheGuardian, "Major breach found in biometrics system used by banks, uk police and defence firms." https://www.theguardian.com/technology/2019/aug/14/major-breach-found-in-biometrics-system-used-by-banks-uk-police-and-defence-firms, 2019.
7. J. M. Such and N. Criado, "Multiparty privacy in social media.," *Commun. ACM*, vol. 61, no. 8, pp. 74–81, 2018.
8. A. Cavoukian, *Evolving FIPPs: Proactive Approaches to Privacy, Not Privacy Paternalism*, pp. 293–309. Dordrecht: Springer Netherlands, 2015.
9. AmnestyInternational, "Surveillance giants: How the business model of google and facebook threatens human right." https://www.amnesty.org/en/documents/pol30/1404/2019/en/, 2019. Accessed 25 Feb. 2020.
10. J. Kaye, E. A. Whitley, D. Lund, M. Morrison, H. J. A. Teare, and K. Melham, "Dynamic consent: a patient interface for twenty-first century research networks," in *European Journal of Human Genetics*, 2015.
11. EuropeanParlamient and of the Council, "Directive 95/46/ec (general data protection regulation)," 2016. Last accessed 25 Feb. 2020.
12. "The california consumer privacy act of 2018, c.c.p.a 1.81.5.), sec 1798.100,." https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375, (2018). Acccessed 7 Apr 2020.
13. K. Fatema, E. Hadziselimovic, H. Pandit, C. Debruyne, D. Lewis, and D. OSullivan, "Compliance through informed consent: Semantic based consent permission and data management model," in *PrivOn@ISWC*, 2017.
14. M. Geko and S. Tjoa, "An ontology capturing the interdependence of the general data protection regulation (GDPR) and information security," in *Proceedings of the Central European Cybersecurity Conference 2018 on - CECC 2018*, ACM Press, 2018.
15. M. Palmirani, M. Martoni, A. Rossi, C. Bartolini, and L. Robaldo, "PrOnto: Privacy ontology for legal reasoning," in *Electronic Government and the Information Systems Perspective*, pp. 139–152, Springer International Publishing, 2018.
16. H. J. Pandit, C. Debruyne, D. O'Sullivan, and D. Lewis, "Gconsent - a consent ontology based on the gdpr," in *The Semantic Web* (P. Hitzler, M. Fernández, K. Janowicz, A. Zaveri, A. J. Gray, V. Lopez, A. Haller, and K. Hammar, eds.), (Cham), pp. 270–282, Springer International Publishing, 2019.
17. B. Knijnenburg, "Privacy? I Cannot Even! making a case for user-tailored privacy," *IEEE Security & Privacy*, vol. 15, no. 4, pp. 62–67, 2017.
18. A. Claesson and T. E. Bjrstad, "Out of control - A review of data sharing by popular mobile apps," tech. rep., Norwegian Consumer Council, 2020.
19. S. Zimmeck, Z. Wang, L. Zou, R. Iyengar, B. Liu, F. Schaub, S. Wilson, N. Sadeh, S. Bellovin, and J. Reidenberg, "Automated analysis of privacy requirements for mobile apps," in *2016 AAAI Fall Symposium Series*, 2016.
20. J. Knackmuss, E. Clausing, and R. Creutzburg, "Investigation of security relevant aspects of android ehealthapps: permissions, storage properties and data transmission," *Electronic Imaging*, vol. 2017, no. 6, pp. 65–75, 2017.

21. B. Friedman, E. Felten, and L. I. Millett, "Informed consent online: A conceptual model and design principles," *University of Washington Computer Science & Engineering Technical Report 00–12–2*, vol. 8, 2000.
22. C. J. Bonnici and L. Coles-Kemp, "Principled electronic consent management: A preliminary research framework," in *2010 International Conference on Emerging Security Technologies*, pp. 119–123, IEEE, 2010.
23. KantaraInitiave, "Kantara-consent receipt specification 2 version," Tech. Rep. Inc., 2018. https://kantarainitiative.org/file-downloads/consent-receipt-specification-v1-1-0/ Last accessed 25. Feb 2020.
24. Kantara, "User-Managed Access (UMA) 101 Eve Maler, Kantara Initiative UMA Work Group chair," 2019.
25. M. Robol, E. Paja, M. Salnitri, and P. Giorgini, "Modeling and reasoning about privacy-consent requirements," in *The Practice of Enterprise Modeling* (R. A. Buchmann, D. Karagiannis, and M. Kirikova, eds.), (Cham), pp. 238–254, Springer International Publishing, 2018.
26. M. C. Mont, V. Sharma, and S. Pearson, "Encore: dynamic consent, policy enforcement and accountable information sharing within and across organisations," tech. rep., 2012.
27. S. Pearson and M. Casassa-Mont, "Sticky policies: An approach for managing privacy across multiple parties," *Computer*, vol. 44, no. 9, pp. 60–68, 2011.
28. S. Pearson, "Privacy, security and trust in cloud computing," in *Privacy and Security for Cloud Computing*, IEEE, 2013.
29. G. Spyra, W. J. Buchanan, and E. Ekonomou, "Blockchain and Git repositories for sticky policies protected OOXML," in *FTC 2017 - Future Technologies Conference 2017*, Vancouver, Canada, 2017.
30. G. Spyra, W. J. Buchanan, and E. Ekonomou, "Sticky policies approach within cloud computing," *Computers & Security*, vol. 70, pp. 366–375, sep 2017.
31. A. Das, M. Degeling, D. Smullen, and N. Sadeh, "Personalized Privacy Assistants for the Internet of Things: Providing Users with Notice and Choice," *IEEE Pervasive Computing*, vol. 17, pp. 35–46, Jul 2018.
32. G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles, "Towards a better understanding of context and context-awareness," in *International symposium on handheld and ubiquitous computing*, pp. 304–307, Springer, 1999.
33. S. Pearson and Y. Shen, "Context-aware privacy design pattern selection," in *Proceedings of the 7th International Conference on Trust, Privacy and Security in Digital Business*, TrustBus10, (Berlin, Heidelberg), p. 6980, Springer-Verlag, 2010.
34. N. F. Noy, D. L. McGuinness, *et al.*, "Ontology development 101: A guide to creating your first ontology. 2001," *See http://protege. stanford. edu/publications*, 2004.
35. F. Loukil, C. Ghedira-Guegan, K. Boukadi, and A. N. Benharkat, "LIoPY: A legal compliant ontology to preserve privacy for the internet of things," in *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, IEEE, jul 2018.
36. M. F. Arruda and R. F. Bulcão Neto, "Toward a lightweight ontology for privacy protection in iot," in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC 19, (New York, NY, USA), p. 880888, Association for Computing Machinery, 2019.
37. M. Gharib, P. Giorgini, and J. Mylopoulos, "Towards an ontology for privacy requirements via a systematic literature review," in *Conceptual Modeling*, pp. 193–208, Springer International Publishing, 2017.

38. D. Haynes, "The nature of risk and the privacy calculus." https://blogs.city.ac.uk/privacycalculus/the-nature-of-risk-in-the-privacy-calculus/methodology/, 2019.
39. M. Li and R. Samavi, "Dsap: Data sharing agreement privacy ontology," in *SWAT4LS*, 2018.
40. W. Xiong and R. Lagerstrm, "Threat modeling a systematic literature review," *Computers & Security*, vol. 84, pp. 53 – 69, 2019.
41. M. Deng, K. Wuyts, R. Scandariato, B. Preneel, and W. Joosen, "A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements," *Requirements Engineering*, vol. 16, pp. 3–32, Mar 2011.
42. K. Wuyts and W. Joosen, "Linddun privacy threat modeling: a tutorial," techreport CW685, Department of Computer Science, KU Leuven, July 2015.
43. C. Perera, C. Liu, R. Ranjan, L. Wang, and A. Y. Zomaya, "Privacy-knowledge modeling for the internet of things: A look back," *Computer*, vol. 49, pp. 60–68, Dec 2016.
44. D. L. Costa, M. J. Albrethsen, M. L. Collins, S. J. Perl, G. J. Silowash, and D. L. Spooner, "An insider threat indicator ontology," Tech. Rep. REPORT, 2016.
45. A. H. Celdran, F. J. G. Clemente, M. G. Perez, and G. M. Perez, "SeCoMan: A semantic-aware policy framework for developing privacy-preserving and context-aware smart applications," *IEEE Systems Journal*, vol. 10, pp. 1111–1124, sep 2016.
46. PrivacyInternational, "No body's business but mine: How menstruation apps are sharing your data." https://privacyinternational.org/long-read/3196/no-bodys-business-mine-how-menstruations-apps-are-sharing-your-data, 2020. Accessed 7 Apr 2020.
47. E. Okoyomon, N. Samarin, P. Wijesekera, A. Elazari Bar On, N. Vallina-Rodriguez, I. Reyes, Á. Feal, and S. Egelman, "On the ridiculousness of notice and consent: Contradictions in app privacy policies," *Workshop on Technologyand Consumer Protection (ConPro)*, 2019.
48. A. Benjamin, M. S. Yaseer, W. Justin, E. William, R. Bradley, S. Kapil, and E. Serge, "Actions speak louder than words: Entity-sensitive privacy policy and data flow analysis with policheck," in *29th USENIX Security Symposium (USENIX Security 20)*, (Boston, MA), USENIX Association, Aug. 2019.
49. WPart19, "Guidelines on transparency under regulation 2016/679," 2016.
50. A. T. der Konferenz der unabhngigen Datenschutzbehrden des Bundes und der Lnder, "Das standard-datenschutzmodell - eine methode zur datenschutzberatung und -prfung auf der basis einheitlicher gewhrleistungsziele, v.2," 2018.
51. M. Louis, "Enisa threat taxonomy: A tool for structuring threat information initial version, 1.0," tech. rep., ENISA, Heraklion, 2016.
52. TheVerge, "Vermont attorney general is suing clearview ai over its controversial facial recognition app." https://www.theverge.com/2020/3/11/21174613/clearview-ai-sued-vermont-attorney-general-facial-recognition-app-database, 2020. Accessed 7 Apr 2020.
53. A. Cavoukian, S. Taylor, and M. E. Abrams, "Privacy by Design: essential for organizational accountability and strong business practices," *Identity in the Information Society*, vol. 3, pp. 405–413, jun 2010.
54. L. Sion, D. Van Landuyt, K. Wuyts, and W. Joosen, "Privacy risk assessment for data subject-aware threat modeling," in *2019 IEEE Security and Privacy Workshops (SPW)*, pp. 64–71, May 2019.